PROJECT PROPOSAL - MACHINE LEARNING

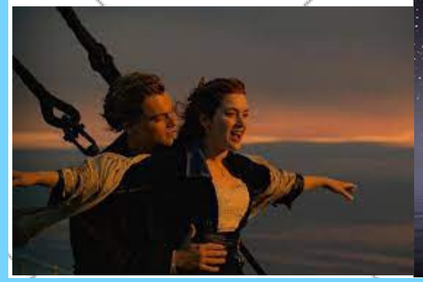
Presented by:
Darius

TODAY'S AGENDA

1. Dataset – Chosen Dataset
2. Dataset – Problem Statement/Goals & Description
3. Dataset – Sample of Chosen Dataset
4. Challenges

CHOSEN DATASET

Titanic



1 Problem Statement / Goal

- To use machine learning to create a model that predicts which passengers survived the Titanic shipwreck
- Predict if a passenger survived the sinking of the Titanic or not.

2 Size of Dataset

- There is a total of 891 rows & 12 columns (Passenger Records)
- There are no duplicate records found in dataset

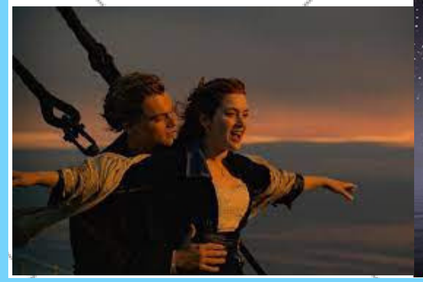
3 Datatypes Found

- int64
- objectve
- float64

Source: https://www.kaggle.com/competitions/titanic/overview

CHOSEN DATASET

Titanic



4

Features found in dataset

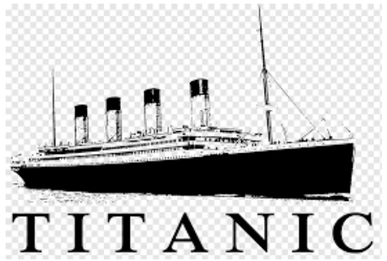
- PassengerID
- PClass (Ticket Class)
- Name
- Sex (Gender)
- Age (In Years)
- SibSp (# of siblings / spouses aboard the Titanic)
- Parch # of parents/ children aboard the Titanic
- Ticket (Ticket Number)
- Fare (Fare Price)
- Cabin (Cabin Number)
- Embarked (Port of Embarkation)

Source: https://www.kaggle.com/competitions/titanic/overview

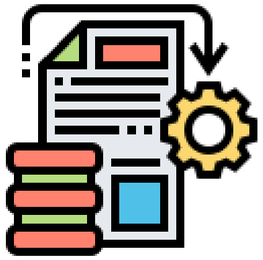
SAMPLE OF DATASET

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
5	897	3	Svensson, Mr. Johan Cervin	male	14.0	0	0	7538	9.2250	NaN	S
6	898	3	Connolly, Miss. Kate	female	30.0	0	0	330972	7.6292	NaN	Q
7	899	2	Caldwell, Mr. Albert Francis	male	26.0	1	1	248738	29.0000	NaN	S
8	900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18.0	0	0	2657	7.2292	NaN	C
9	901	3	Davies, Mr. John Samuel	male	21.0	2	0	A/4 48871	24.1500	NaN	S
10	902	3	Ilieff, Mr. Ylio	male	NaN	0	0	349220	7.8958	NaN	S
11	903	1	Jones, Mr. Charles Cresson	male	46.0	0	0	694	26.0000	NaN	S
12	904	1	Snyder, Mrs. John Pillsbury (Nelle Stevenson)	female	23.0	1	0	21228	82.2667	B45	S
13	905	2	Howard, Mr. Benjamin	male	63.0	1	0	24065	26.0000	NaN	S
14	906	1	Chaffee, Mrs. Herbert Fuller (Carrie Constance...	female	47.0	1	0	W.E.P. 5734	61.1750	E31	S
15	907	2	del Carlo, Mrs. Sebastiano (Argenia Genovesi)	female	24.0	1	0	SC/PARIS 2167	27.7208	NaN	C
16	908	2	Keane, Mr. Daniel	male	35.0	0	0	233734	12.3500	NaN	Q
17	909	3	Assaf, Mr. Gerios	male	21.0	0	0	2692	7.2250	NaN	C
18	910	3	Ilmakangas, Miss. Ida Livija	female	27.0	1	0	STON/O2. 3101270	7.9250	NaN	S
19	911	3	Assaf Khalil, Mrs. Mariana (Miriam")"	female	45.0	0	0	2696	7.2250	NaN	C
20	912	1	Rothschild, Mr. Martin	male	55.0	1	0	PC 17603	59.4000	NaN	C
21	913	3	Olsen, Master. Artur Karl	male	9.0	0	1	C 17368	3.1708	NaN	S
22	914	1	Flegenheim, Mrs. Alfred (Antoinette)	female	NaN	0	0	PC 17598	31.6833	NaN	S
23	915	1	Williams, Mr. Richard Norris II	male	21.0	0	1	PC 17597	61.3792	NaN	C
24	916	1	Ryerson, Mrs. Arthur Larned (Emily Maria Borie)	female	48.0	1	3	PC 17608	262.3750	B57 B59 B63 B66	C
25	917	3	Robins, Mr. Alexander A	male	50.0	1	0	A/5. 3337	14.5000	NaN	S
26	918	1	Ostby, Miss. Helene Ragnhild	female	22.0	0	1	113509	61.9792	B36	C
27	919	3	Daher, Mr. Shedid	male	22.5	0	0	2698	7.2250	NaN	C
28	920	1	Brady, Mr. John Bertram	male	41.0	0	0	113054	30.5000	A21	S
29	921	3	Samaan, Mr. Elias	male	NaN	2	0	2662	21.6792	NaN	C
30	922	2	Louch, Mr. Charles Alexander	male	50.0	1	0	SC/AH 3085	26.0000	NaN	S

WORKFLOW TO TACKLE THE PROBLEM

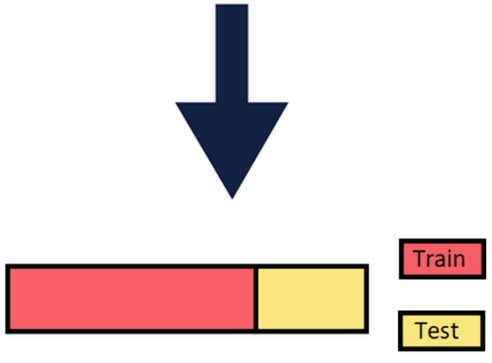


Data

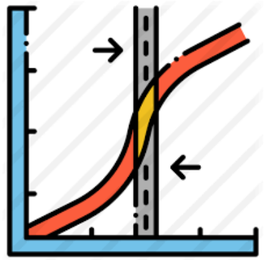


Data Pre-processing

Data Analysis



Train Test Split



Logistic Regression Model

Evaluation

CHALLENGES

MISSING VALUES

```
print(df_titanic.isnull().sum())
#missing data on age, cabin & embarked
```

[7] ✓ 0.7s

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

STATISTIC KNOWLEDGE

- Being weak in statistics, I have difficulty differentiating what type of model to use.


THANK
YOU!
Canva